Spring 2-17-2011

Identifying Horizontal Gene Transfer Using Anomalies In Protein Structures And Sequences

Venkat Ram B. Santosh
venkatrambs@gmail.com

IDENTIFYING HORIZONTAL GENE TRANSFER USING ANOMALIES IN PROTEIN STRUCTURES AND SEQUENCES

By

Venkat Ram B. Santosh

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Peter Revesz and Professor Mark A. Griep

Lincoln, Nebraska

January, 2011

IDENTIFYING HORIZONTAL GENE TRANSFER USING ANOMALIES IN PROTEIN STRUCTURES AND SEQUENCES

Venkat Ram Santosh, M.S.

University of Nebraska, 2011

Advisor: Peter Revesz and Mark A. Griep

Genetics has traditionally focused on *vertical gene transfer*, which is the passing of the genetic material of an organism to its offspring. However, recent studies in genetics increased the awareness that *horizontal gene transfer*, which is the passing of the genetic material of an organism to another organism that is not its offspring, is also a significant phenomenon. Horizontal gene transfer is thought to play a major role in the natural evolution of bacteria, such as, when several different types of bacteria all suddenly develop the same drug resistance genes. Artificial horizontal gene transfer occurs in genetic engineering.

This thesis provides methods to detect horizontal gene transfer among bacteria using BLAST and DaliLite measures of protein sequence and structural similarities. This research is novel and unique because no previous horizontal gene transfer study worked directly on protein sequences and structures. The main method is a computer algorithm to detect horizontal gene transfer among different COG classifications of proteins. The thesis also considers visual structural comparisons and sequence alignments using the 'Jmol' tool. Finally, the thesis considers the possibility that the method yields false positives.

Acknowledgements

I would like to express my sincere gratitude to Prof. Peter Revesz, my advisor from the Department of Computer Science and Engineering, for accepting me as his student and introducing me to bioinformatics and the study of horizontal gene transfer. His constant guidance, ideas and support throughout the completion of my work were the key for my thesis. I owe a special thanks to Prof. Mark Griep, my co-advisor from the Department of Chemistry, for his willingness to always help taking time from his busy schedule as the Vice Chair of the Department of Chemistry. I would like to thank also Prof. Jitender Deogun for agreeing to serve on my committee and for providing valuable feedback and comments on my thesis.

I would like to thank also several students in our research group. Swetha Billa provided support, motivation and help throughout the thesis. Thomas Triplet and Matt Shortridge answered patiently my innumerable questions when I started this work. Derek Weitzel helped run DaliLite on the grid in the Schorr center.

I would like to thank all my friends in Lincoln, especially my close friend Rahul Maru for being supportive and helping me out during the toughest of times.

Finally, I would like to thank my parents. Without their encouragement and support this thesis could not have been possible.

Contents

List of Figures

List of Tables

List of Abbreviations

1 Introduction ……... 1

1.1 Horizontal Gene Transfer …… 1

2 Background ……... 3

2.1 Basic Concepts of Biology ……..... 3

2.1.1 Bacteria ……. 3

2.1.1.1 Classification of Bacteria based on Phyla ……..... 4

2.1.1.2 Gram Staining ……..... 5

2.1.2 Amino Acids ……. 6

2.1.3 Proteins ……..... 6

2.1.4 Codons …….. 7

2.2 Basics of HGT …….. 8

2.2.1 Mechanisms of HGT ……. 8

2.2.2 Examples of Possible HGT …….. 9

2.3 Related Study ……. 11

2.3.1 Biological Databases Used ……...……. 11

2.3.1.1 Protein Data Bank (PDB) …… 11

2.3.1.2 Clusters of Orthologous Groups (COG) Database ……... 11

2.3.1.3 Gene Ontology (GO) Database …… 12

2.3.1.4 PROFESS (PROtein Function, Evolution, Structure and Sequence) Database ……. 12

2.3.2 Biological Tools Used ……. 13

2.3.2.1 BLAST ……..…… 13

2.3.2.2 DaliLite …… 14

2.3.2.3 Jmol ……. 16

2.3.3 Methods Currently Used to Detect HGT ……... 17

2.3.3.1 Phylogeny-Based Detection of HGT ...…… 17

2.3.3.2 Distance-Based Detection of HGT ...……..……. 18

2.3.3.3 Composition-Based Detection of HGT …… 19

3 Methodology …… 21

3.1 The Method …… 22

3.2 Automation …… 24

4 Analysis and Results ……... 27

4.1 Summary of Suspected HGT ...…….. 36

4.2 Detailed Analysis of COG-503 ...……... 37

4.3 Detailed Analysis of COG-596 ...……... 41

4.4 Detailed Analysis of COG-604 ...……... 44

4.5 Detailed Analysis of COG-1278 ...……. 48

4.6 False Positives ...…… 51

5 Relative COG Functional Similarity Based on GO terms ...……... 52

5.1 Introduction …… 52

5.1.1 Hamiltonian Distance 53

5.2 Method 54

5.2.1 Method 1 54

5.2.2 Method 2 ……. 55

5.3 Results …… 56

6 Conclusion and Future work ……. 60

6.1 Conclusion ……. 60

6.2 Future Work ...…… 61

References

List of Figures

2.1 Cell wall comparison of a Gram-positive and Gram-negative bacterium ……... 5

3.1 Database schema used for automation ……... 25

3.2 Flow chart of the method ……... 26

4.1 Pre-calculated jFATCAT-rigid structure alignment results 2DY0 (*E. coli*) vs. 1A98 (*E. coli*) …… 39

4.2 Pre-calculated jFATCAT-rigid structure alignment results 2DY0 (*E. coli*) vs. 1O57 (*Bacillus subtilis*) ……. 39

4.3 Sequence alignment results 2DY0 (*E. coli*) vs. 1A98 (*E. coli*) ……. 40

4.4 Sequence alignment results 2DY0 (*E. coli*) vs. 1O57 (*Bacillus subtilis*) …….. 40

4.5 Pre-calculated jFATCAT-rigid structure alignment results 1M33 (*E. coli*) vs. 1U2E (*E. coli*) …… 42

4.6 Pre-calculated jFATCAT-rigid structure alignment results 1M33 (*E. coli*) vs. 1WOM (*Bacillus subtilis*) ……. 42

4.7 Sequence alignment results 1M33 (*E. coli*) vs. 1U2E (*E. coli*) ……. 43

4.8 Sequence alignment results 1M33 (*E. coli*) vs. 1WOM (*Bacillus subtilis*) ……... 43

4.9 Pre-calculated jFATCAT-rigid structure alignment results 1O89 (*E. coli*) vs. 1QOR (*E. coli*) …… 45

4.10 Pre-calculated jFATCAT-rigid structure alignment results 1O89 (*E. coli*) vs. 1TT7 (*Bacillus subtilis*) ……. 45

4.11 Sequence alignment results 1O89 (*E. coli*) vs. 1QOR (*E. coli*) ……... 46

4.12 Sequence alignment results 1O89 (*E. coli*) vs. 1TT7 (*Bacillus subtilis*) ……. 47

4.13 Pre-calculated jFATCAT-rigid structure alignment results 3MEF (*E. coli*) vs. 2BH8 (*E. coli*) …… 49

4.14 Pre-calculated jFATCAT-rigid structure alignment results 3MEF (*E. coli*) vs. 2ES2 (*Bacillus subtilis*) ……. 49

4.15 Sequence alignment results 3MEF (*E. coli*) vs. 2BH8 (*E. coli*) …….. 50

4.16 Sequence alignment results 3MEF (*E. coli*) vs. 2ES2 (*Bacillus subtilis*) …… 50

List of Tables

3.1 Example of Documented Data ……... 23

4.1 Z-score Structural comparison between *Escherichia coli* and *Bacillus subtilis* …… 29

4.2 Z-score Structural comparison between *Escherichia coli* and *Staphylococcus aureus* ……... 30

4.3 Z-score Structural comparison between *Escherichia coli* and *Bacillus stearothermophilus* …...……... 31

4.4 Z-score Structural comparison between *Escherichia coli* and *Streptococcus pneumonia* …… 32

4.5 Z-score Structural comparison between *Escherichia coli* and *Lactococcus lactis* 32

4.6 Z-score Structural comparison between *Escherichia coli* and *Bacillus anthracis* … 33

4.7 Z-score Structural comparison between *Escherichia coli* and *Bacillus megaterium* 33

4.8 Summary of candidates for HGT among the compared protein structures …….. 34

4.9 Summary of Proteins suspected as HGT ……... 36

4.10 COG- 503 in Comparison between *Escherichia coli* and *Bacillus subtilis* ……..... 38

4.11 COG- 596 in Comparison between *Escherichia coli* and *Bacillus subtilis* ……..... 41

4.12 COG-604 in Comparison between *Escherichia coli* and *Bacillus subtilis* …….. 44

4.13 COG-1278 in Comparison between *Escherichia coli* and *Bacillus subtilis* ...……. 48

5.1 Results of COG functional similarity with Method 1 and Method 2 ……. 56

List of Abbreviations

HGT- Horizontal Gene Transfer

LGT- Lateral Gene Transfer

DNA- Deoxyribonucleic acid

RNA- Ribonucleic acid

PDB- Protein Data Bank

COG- Clusters of Orthologous Groups

GO- Gene Ontology

NCBI- National Center for Biotechnology Information

PROFESS- PROtein Functions, Evolution, Structures and Sequences Database

BLAST- Basic Local Alignment Search Tool

Chapter 1

Introduction

1.1 Horizontal Gene Transfer

Horizontal gene transfer (HGT), which is also called *lateral gene transfer*, is any process in which an organism incorporates genetic material from another organism without being the offspring of that organism. In contrast, *vertical gene transfer* occurs when an organism receives genetic material from its ancestor, e.g. its parent or a species from which it evolved. Genetics traditionally focused on vertical gene transfer, but there is a growing awareness that horizontal gene transfer is also a highly significant phenomenon, and among single-celled organisms perhaps the dominant form of genetic transfer. During HGT, genetic material can pass between organisms that need not be of the same species, genus, sub-kingdom or even kingdom of life.

Detection of HGT is complicated and difficult. Horizontal gene transfer was first described in Japan in a 1959 publication that demonstrated the transfer of antibiotic resistance between different species of bacteria.[23][1] In the late 1980s while conducting research on some newly sequenced bacterial and archaeal gene families, scientists began

to notice that some genetic information was common among them, some bacteria possessed the archaeal type of an enzyme, and some of the archaea contained the bacterial versions. These discoveries were shaking the original metaphor of “the tree of life.”[25]

Increasing studies of genes and genomes are indicating that considerable horizontal transfer has occurred between prokaryotes. The phenomenon appears to have had some significance for unicellular eukaryotes as well. There is some evidence that even higher plants and animals have been affected and this has raised concerns for safety.[22] Due to the increasing amount of evidence suggesting the importance of these phenomena for evolution, molecular biologists have described horizontal gene transfer as a “new paradigm for biology”. It should also be noted that the process may be a hidden hazard of genetic engineering, as it may allow dangerous transgenic DNA (which is optimized for transfer) to spread from species to species.

Chapter 2

Background

2.1 Basic Concepts of Biology

2.1.1 Bacteria

Bacteria are unicellular prokaryote microorganisms. Their size typically ranges a few millimeters in length and they vary in shapes from sphere to rod to spiral. The cell of a bacterium is surrounded by a cell membrane which encloses the contents of the cell and helps hold the nutrients, proteins and other essential components of the cytoplasm within the cell. They do not have membrane bound organelles in their cytoplasm, so do not have a their nucleus, mitochondria or chloroplasts.[3]

Bacteria were first observed in 1676, which means they have been under study since a very long time. Horizontal gene transfer in bacteria is thought to be a significant cause of increased drug resistance; when one bacterial cell acquires resistance, it can quickly transfer the resistance genes to many other bacterial species.[4] Enteric bacteria appear to exchange genetic material with each other within the gut in which they live.

2.1.1.1 Classification of Bacteria based on phyla

Based on the morphology, DNA sequencing, conditions required and biochemistry, scientists classify bacteria into the following *phyla*:

1. Aquificae
2. Xenobacteria
3. Fibrobacter
4. Bacteroids
5. Firmicutes
6. Planctomycetes
7. Chrysogenetic
8. Cyanobacteria
9. Thermomicrobia
10. Chlorobia
11. Proteobacteria
12. Spirochaetes
13. Flavobacteria
14. Fusobacteria
15. Verrucomicrobia

The main two phylum classifications of bacteria that we are going to be interested in are *Firmicutes* and *Proteobacteria*. *Firmicutes* belongs to *Gram positive* bacteria and *Proteobacteria* belongs to *Gram negative* bacteria.

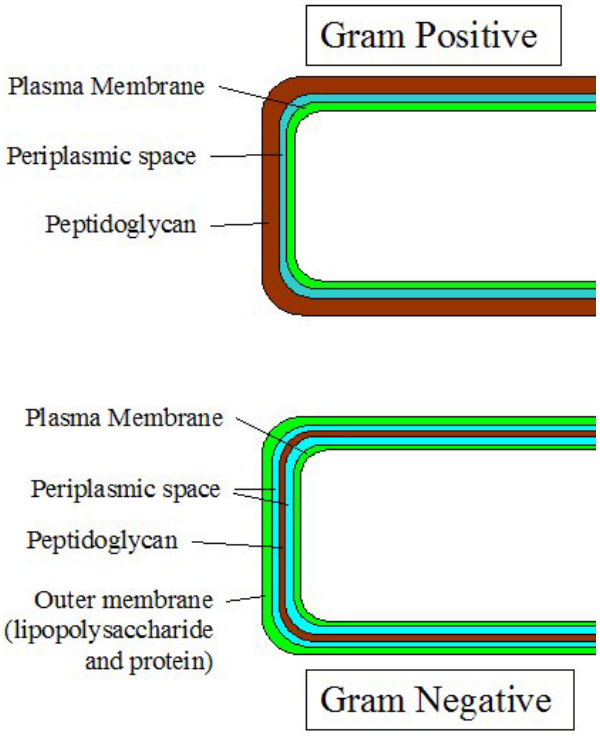


Figure 2.1: Cell wall comparison of a Gram-positive and Gram-negative bacterium [20]

2.1.1.2 Gram Staining

Gram staining by a crystal violet dye is generally the first step used to identify a bacterium. Gram staining differentiates bacteria based on the chemical and physical properties of the cell walls into the following two different types: [5]

Gram negative: Bacteria that do not retain the crystal violet dye.

Gram positive: Bacteria that retain the crystal violet dye. [30]

2.1.2 Amino Acids

Amino acids play a major role as building blocks of proteins. Every protein is chemically defined by the order of amino acid residues and their primary structure, which in turn influences the secondary structure, tertiary structure or quaternary structure. Similar to letters of a language that can be combined in different combination to form words, amino acids combine in various stable combination of varying length to form a vast variety of proteins. There are 20 amino acids that are found within proteins. It is important to know the structure and chemistry of amino acids to understand the proteins, enzymes and nucleic acids.[9]

2.1.3 Proteins

Proteins are organic compounds made of amino acids arranged in a linear chain and folded. The amino acids in a polymer are joined together by the peptide bonds between the carboxyl and amino groups. The sequence of the gene decides the sequence of amino acids in a protein.[29]

Discovering the structure of the protein can provide insight into the function that the protein performs. So understanding and comparing protein structures is crucial. Most proteins fold into unique 3-dimensional structures. The shape into which a protein naturally folds is known as its native conformation. There are four distinct aspects of a protein's structure:

1. Primary structure is the amino acid sequence.
2. Secondary structure is the regularly repeating local structures stabilized by hydrogen bonds.

3. Tertiary structure is the overall shape of a single protein molecule; the spatial relationship of the secondary structures to one another.
4. Quaternary structure is the structure formed by several protein molecules (polypeptide chains).

2.1.4 Codons

The tri-nucleotide sequences are called codons. The genetic code defines a mapping between the codons and amino acids.

Start codon: Translation starts with a chain initiation codon (start codon). The most common start codon is AUG.

Stop codon: It is a nucleodtide triplet within messenger RNA that signals a termination of translation.[13]

The several stop codons are as follows:

- in RNA:
  - UAG
  - UAA
  - UGA
- in DNA:
  - TAG
  - TAA
  - TGA

2.2 Basics of HGT

2.2.1 Mechanisms of HGT

Horizontal gene transfer could occur by several mechanisms between organisms. There are three basic mechanisms as described below.

- **Transformation -** The uptake of naked DNA is a common mode of horizontal gene transfer that can mediate the exchange of any part of a chromosome; this process is most common in bacteria that are naturally transformable; typically only short DNA fragments are exchanged.
- **Conjugation** - The transfer of DNA mediated by conjugal plasmids or conjugal transposons; requires cell to cell contact but can occur between distantly related bacteria or even bacteria and eukaryotic cells; can transfer long fragments of DNA.
- **Transduction** - The transfer of DNA by phage requires that the donor and recipient share cell surface receptors for phage binding and thus is usually limited to closely related bacteria; the length of DNA transferred is limited by the size of the phage head.
- Gene transfer agents, virus-like elements encoded by the host that are found in the alphaproteobacteria order Rhodobacterales.[21]

Each of these methods of genetic exchange can introduce sequences of DNA that share little homology with the remaining DNA of the recipient cell. If there are homologous sequences shared between the donor DNA and the recipient chromosome, the donor sequences can be stably incorporated into the recipient chromosome by genetic

recombination. If the homologous sequences flank sequences that are absent in the recipient, the recipient may acquire an insertion from another strain of unrelated bacteria. Such insertions can be small or quite large. Large insertions that have been acquired from another bacterium (often inferred from differences in GC content or codon usage) and are absent from related strains of bacteria are called "islands".

2.2.2 Examples of HGT

In Viruses

The virus called *Mimivirus* can itself be infected by a virus called *Sputnik*. "Sputnik's genome reveals further insight into its biology. Although 13 of its genes show little similarity to any other known genes, three are closely related to mimivirus and mamavirus genes, perhaps cannibalized by the tiny virus as it packaged up particles sometime in its history. This suggests that the satellite virus could perform horizontal gene transfer between viruses — paralleling the way that bacteriophages ferry genes between bacteria."[19][24]

In Prokaryotes

Horizontal gene transfer is common among bacteria, even very distantly-related ones. This process is thought to be a significant cause of increased drug resistance; when one bacterial cell acquires resistance, it can quickly transfer the resistance genes to many species. Enteric bacteria appear to exchange genetic material with each other within the gut in which they live.[12]

In Eukaryotes

Analysis of DNA sequences suggests that horizontal gene transfer has also occurred within eukaryotes, from their chloroplast and mitochondrial genome to their nuclear genome. As stated in the endosymbiotic theory, chloroplasts and mitochondria probably originated as bacterial endosymbionts of a progenitor to the eukaryotic cell.

Horizontal transfer of genes from bacteria to some fungi, especially the yeast *Saccharomyces cerevisiae*, has been documented.

There is also recent evidence that the azuki bean beetle has somehow acquired genetic material from its (non-beneficial) endosymbiont *Wolbachia*.[18]

Traditionally only the Vertical gene transfers i.e. flow of genes from parent to child was considered for all the study. Given two distantly related bacteria that have exchanged a gene, a phylogenetic tree including those species will show them to be closely related because that particular gene is the same, even though most other genes are dissimilar.

The most common gene that is used for constructing phylogenetic relationships in prokaryotes is the ‘16s rRNA’ gene. But recent study shows 16s rRNA genes can also be horizontally transferred. So the validity of 16s rRNA-constructed phylogenetic trees must be reevaluated.

2.3 Related Study

2.3.1 Biological Databases Used

2.3.1.1 Protein Data Bank (PDB)[6]

It is a single worldwide repository of information about 3D structures of large biological molecules, including proteins and nucleic acids. The data on PDB is available for free, these data are actually submitted by biologists and biochemists from around the world which is first reviewed and then published. The biologists and biochemists typically obtain their data by X-ray crystallography or NMR spectroscopy. The PDB is a key resource in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding agencies, such as the NIH in the USA, now require scientists to submit their structure data to the PDB.

Understanding the shape of the molecule is important to understand the functioning of the molecule. Thus, PDB helps the scientific world in determining the structure's role in human health and disease and also in drug development. The PDB database can be found here: *http://www.pdb.org/*

2.3.1.2 Clusters of Orthologous Groups (COG) Database[33]

COG is a classification of proteins generated by comparing the protein sequences of complete genomes. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

The National Center for Bio-technology Information (NCBI) advances science and health by providing access to biomedical and genomic information. The NCBI maintains and

updates the COG database. There are currently 66 Unicellular organisms consisting of 138458 proteins The proteins form 4873 COGs. The COG database can be found at *http://www.ncbi.nlm.nih.gov/COG*

2.3.1.3 Gene Ontology (GO) Database[7]

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. GO is not a database of gene sequences, nor a catalog of gene products. Rather, GO describes how gene products behave in a cellular context. The GO Database can be found at *http://www.geneontology.org/*

2.3.1.4 PROFESS (PROtein Function, Evolution, Structure and Sequence) Database[35]

PROFESS is a genome biology database system, developed at University of Nebraska-Lincoln, to assist in the functional and evolutionary analysis of the proteins. Fourteen sources of data were integrated to create PROFESS using a local-as-view (LAV) modular approach.

There are about 1100 molecular biology databases freely available to the public online and there is no proper integration between these databases. To address more complex question, biologist are often required to design their own databases and have to put together the data from these various biological databases. The PROFESS database integrates these diverse biological databases under a single platform. This unique integration of various databases makes the profess database of great use to this research.

One of them is the COG-PDB id relation or mapping that PROFESS has created was used in this research.

The PROFESS Database can be found at *http://cse.unl.edu/~profess/*

2.3.2 Biological Tools Used

2.3.2.1 BLAST[2]

Basic Local Alignment Search Tool (BLAST) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. BLAST helps researches to compare a sequence with another sequence or a database of sequences and identify the sequences that are similar to the query sequence above a specified threshold. Different types of BLASTs are available according to the query sequences. BLAST takes the FASTA or GENBank format of the amino acid sequence as input. BLAST output can be delivered in a variety of formats. These formats include HTML, plain text, and XML formatting.

BLAST is actually a family of programs (all included in the blastall executable). These include, Nucleotide-nucleotide BLAST (blastn): This program, given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies.

Protein-protein BLAST (blastp): This program, given a protein query, returns the most similar protein sequences from the protein database that the user specifies.

Position-Specific Iterative BLAST (PSI-BLAST): This program is used to find distant relatives of a protein. PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than a standard protein-protein BLAST.

Nucleotide 6-frame translation-protein (blastx): This program compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx): It translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database. The purpose of tblastx is to find very distant relationships between nucleotide sequences.

Protein-nucleotide 6-frame translation (tblastn): This program compares a protein query against the all six reading frames of a nucleotide sequence database.

Large numbers of query sequences (megablast): It concatenates many input sequences together to form a large sequence before searching the BLAST database, then post-analyze the search results to extract individual alignments and statistical values.

Of all these programs, BLASTn and BLASTp are the most commonly used because they use direct comparisons and do not require translations.

2.3.2.2 DaliLite[10][14]

DaliLite is a program for pair wise structural comparison and for structural database searching. It is the standalone version of the popular Dali server search engine. DaliLite can be implemented with a web interface were we can view the similarity z-scores as well as the structures. It can be implemented without the web interface as well. DaliLite

has been ported to Linux and Iris operation system and can be complied in other UNIX based operation system as well.

The Dali method (Distance matrix alignment) is based on a sensitive approach that measures the weighted sum of similarities of intermolecular distances. The Dali server is often used compare newly solved structures to those in the Protein data bank to compare the predicted structure with the actual structure. But the Dali server is accessible only through the network and is too complex and large to run locally. So we use DaliLite which is a standalone program that researchers can run locally and thus helps to compare a large number of structures with ease and the flexibility of running locally is an addition. For the user, DaliLite seems to be taking two PDB ids and giving out the structural comparison results. Internally, it takes two sets of atomic coordinates in PDB format. A visualization script named DaliQuiz converts FSSP alignments to graphical output. The program outputs the FSSP file and displays the Z-score (the structural similarity score) which are normalized with respect to domain size. The Z-score output of DaliLite is of most interest to us in this research.

The Z-score is the measure of the quality of alignment. Z-score above 20 means generally homologous, Z-score between 8-20 means probably homologous and Z-scores less than 2 are not significant.

Dali uses Branch and Bound Search to find the optimal alignment and Monte Carlo Optimization Algorithm to optimize the alignment.

The similarity score is give by the formula.

$$S = \sum_{i \in core} \sum_{j \in core} \left(\theta - \Delta\left(d_{ij}^{A}, d_{ij}^{B}\right)\right) \omega\left(d_{ij}^{A}, d_{ij}^{B}\right)$$

In the formula:

- core is the set of structurally equivalent residue pairs between proteins A and B.
- Δ is the deviation of the intermolecular Cα-Cα intermolecular distance between (i^A,j^A) and (i^B,j^B), relative to their arithmetic mean d.
- θ is the similarity threshold, set empirically to 0.2
- ω is the envelope function and $\omega = \exp(-d^2/r^2)$, where r = 20Å.
- High score means good fit.

And then the Z-score is statistically calculated as

$$Z = \frac{x - \mu}{\sigma}$$

- X is the raw score to be standardized.
- σ is the standard deviation.
- μ is the mean.
- Score < 2.0 are structurally dissimilar.

2.3.2.3 Jmol[17]

Jmol is an open-source java viewer for chemical structures in 3-D. Jmol tool is of great use for students, educators and researchers in chemistry and biochemistry. Since it's an

open source it is available for free, runs on Windows, Mac OS X, Linux and Unix systems. There is a standalone Jmol application and it also has a development tool kit which helps in integration of Jmol with other Java applications. One of the notable features is the Jmol applet that can be embedded in a webpage. Many websites now have the Jmol applet embedded to their site. The Protein data bank website http://www.rcsb.org/ has the Jmol applet embedded for compassion of protein structures.

Jmol supports a wide range of molecular file formats, including Protein Data Bank (pdb), Crystallographic Information File (cif), MDL Molfile (mol), and Chemical Markup Language (CML).

2.3.3 Methods currently Used to Detect HGT

During the past decade, different approaches have been proposed for the detection of HGT, which can be classified in two major categories: (a) the phylogeny-based methods and (b) the composition-based methods. Some of them are described here which helps us understand the uniqueness of the new approach which uses protein structures to detect HGT.

2.3.3.1 Phylogeny-Based Detection of HGT

Phylogeny-based detection of HGT is one of the most commonly used approaches for detecting HGT. It is based on the fact that HGT causes discrepancies in the gene tree as well as create conflict with the species phylogeny. So the methods that use this approach

would compare the gene and species tress which would come up with a set of HGT events to explain the discrepancies among these trees.[34]

When HGT occurs, the evolutionary history of the gene would not agree with the species phylogeny. The gene trees get reconstructed and their disagreements are used to estimate how many events of HGT could have occurred and the donors and recipients of the gene transfer.

Some of the issues when using this method for HGT detection are, determining if the discrepancy is actually a HGT and uniquely identifying the HGT scenario.

The Phylogenetic trees are only partially known and they are reconstructed using Phylogeny reconstruction techniques. The quality of this reconstruction which is usually done statistically has an impact on the HGT detection and sometimes could underestimate or overestimate the number HGT events.

Eliminating these statistical errors is possible but this will lead to non-binary Phylogenetic trees. But this method works with Binary Phylogenetic trees only. So this method will need to be modified to accommodate non-Binary Phylogenetic trees as well.

2.3.3.2 Distance-Based Detection of HGT

The Distance-Based method incorporates *distances* typically used in the Phylogeny-based detection of HGT rather than the trees themselves.[36] This method has many of the strengths of Phylogenetic approaches but avoids some of their pitfalls.

This method uses only the pair-wise distance instead of building the whole trees as in the Phylogeny-based approach, which makes the distance-based approach run much more quickly, allowing scanning of whole genomes. As there is no ‘consensus’ tree in this method, it does not suffer in the cases where no tree matches all of the given data. Instead it just compares the pair-wise distance between species and thus called the Distance-Based method for detecting HGT.

2.3.3.3 Composition-Based Detection of HGT

Although the Phylogeny-Based detection methods are more powerful than the Composition-based methods, especially when the donor is closely related to the recipient genome, they are very time consuming.[8]

The four methodologies commonly employed by Composition-based methods to detect HGT are based on

- The codon adaptation index, codon usage, and GC percentage.(CAI/GC)
- The distributional profile
- The Bayesian model
- The first-order Markov model

All these methods attempt to identify genes with anomalous compositions.

The genomic DNA of different organisms has a particular mean G+C content. Genes in a given genome use the same coding strategy for choices among synonymous codons. That

is, the bias in codon usage is species specific. Statistical methods have been developed to use these anomalies in the GC content to detect HGT.[12]

One notable problem with the compositional approaches is that the codon usage and GC content give different results, each detecting a different set of possible horizontal gene transfers that really didn't match with each other.

Study on these methods show that both the Bayesian models and the Markov models do a good job in detecting HGT when closely related species are studied, though the Markov model is more effective. The CAI/GC method appears to be a less effective approach in the detection of HGT but was very effective in detecting HGT when the foreign genes were from a phylogenetically distant species. The distribution profile method exhibited an average detection level of approximately 50% for foreign genes but failed to go beyond 80% threshold of detection.

If a compositional method with an accurate detection level of horizontally transferred genes can be developed, it could avoid the application of exhaustive processes and slow Phylogenetic reconstructions used in the phylogeny-based approach.

Chapter 3

Methodology

Instead of using the traditional methods for identifying HGT, we devised a novel protein structure-based method (HGT-SBM). When a protein is acquired by HGT, the structure of the protein remains fairly similar to that of the donor organism as it tries to retain close similarities to the function of the donor protein.

We used the COG classification of protein function to look for protein structure anomalies. All proteins under the same COG classification are supposed to have similar function, which evolutionary theory indicates they should have similar structures.

For this research, we try to identify HGT between the bacterial phyla *Firmicutes* and *Proteobacteria*. Most medically important bacteria fall into these two phyla, which diverged hundreds of millions of years ago. During their subsequent evolutions, the proteins in all *Firmicutes* bacteria acquired random mutation but still remained more similar to the other *Firmicutes* bacterial proteins than to the *Proteobacteria* bacterial proteins and vice-versa. Hence any anomalous proteins (i.e. proteins having

characteristics of the other phyla's protein) in either of the phylum would be a very good candidate for a horizontal gene transfer that occurred fairly recently.

3.1 The Method

We chose *E. coli* from *Proteobacteria* and *Bacillus subtilis* from *Firmicutes* as candidates from the two phyla as they have the most number of studied structures from their respective phyla.

PROFESS database was used to get the list of proteins that have been studied in *E. coli* and *Bacillus subtilis* and also the COGs to which they belong to. The COG classification enables us to identify the proteins which are functionally similar.

DaliLite program was used for the structure comparison of the proteins. We first determine the extent of structural similarity of all the proteins in a particular COG within each organism chosen from the two phyla (in this case *E. coli* and *Bacillus subtilis*). Then a pair-wise structural comparison of the proteins between the two organism in each COG is done (in this case the *E. coli* proteins are compared with the *Bacillus subtilis* proteins).

We have approximately 3264 unique PDB IDs in *E. coli* and 494 unique PDB IDs in *Bacillus subtilis*. There are about 88 COGs common in both these organisms. This would result in n * (n-1)/2 pairs of PDB IDs for each COG, where n is the number of proteins in a COG. For the pair-wise comparison between the two organisms within the same COG the number of pairs of PDB IDs would be the cross product of number of proteins in that COG in each organism. For all these cases the averages of the Z-scores for all the pair wise comparison within a COG (for all the common COGs) are documented in a table.

DaliLite gives different Z-scores values for a pair of proteins corresponding to different alignments. We use the best alignment i.e. the highest Z-score value that DaliLite outputs for a given pair. A normalization process is done on the documented Z-score. This is done by choosing the maximum of the 3 average Z-scores values obtained for a COG (one average Z-score from the comparison of proteins within the Proteobacteria, one average Z-score from the comparison of proteins within the Firmicutes and one average Z-score from the comparison of proteins between Proteobacteria & Firmicutes.) All the 3 average Z-scores for a COG are divided by this max Z-score value.

Now we try to compare and look for Z-score anomalies as this will identify protein structure anomalies. Usually the average values of Z-scores for proteins within a COG for both the organisms in comparison are supposed to be pretty similar. So we try to identify those COGs which have high average Z-scores in one organism and a low average Z-score in the other organism. A threshold of 75% was chosen for the average Z-score values to identify as an anomaly. For example of the documented average Z-scores looks like the table below.

Table 3.1: Example of Documented Data

Common COG	*E. coli*	*Bacillus subtilis*	Comparison Z-Score	*E. coli* normalized Z-Score	*Bacillus subtilis* normalized Z-score	Comparison Z-Score normalized
500	11	39.5	15.4	0.27848	1	0.38987

In the above example, the COG 500 is identified as having an anomaly because the average Z-score in COG 500 in *E. coli* is only 11 which is 27.8% of the average Z-score in *Bacillus subtilis* which is 39.5.

Most of the times the reason this happens is, there are one or more proteins in the COG that have dissimilar protein structures compared to the other proteins in the same COG. These proteins are candidates for HGT. Not all anomalous protein structures can be identified as HGT. A careful and a systematic hand curation of the Z-scores must be done to identify or eliminate different PDB structures for the same protein, some of them bound to ligands and some with different conformation. It is also necessary to examine enzyme names to ensure the PDBs are for different proteins with the same COG. Finally, it was necessary to compare super imposed structures to verify that HGT had occurred.

3.2 Automation

When we would like to study many organisms and search for proteins that are candidates for HGT the process should be automated to the extent possible.

First, the data set was assembled by downloading all the protein structures in all the bacteria in the *Firmicutes* and *Proteobacteria* phyla along with their COG classification from the PROFESS database.

The protein structural comparisons were automated such that DaliLite was run on the grid in Holland computing center in Schorr center at UNL. The Z-scores results from DaliLite for all the possible combinations of proteins with in a COG classification were stored,

which gave about 14 million lines of output Z-scores corresponding to all the combinations. All these data was converted into a small database for ease to program and use the data.

The database scheme looks like below:

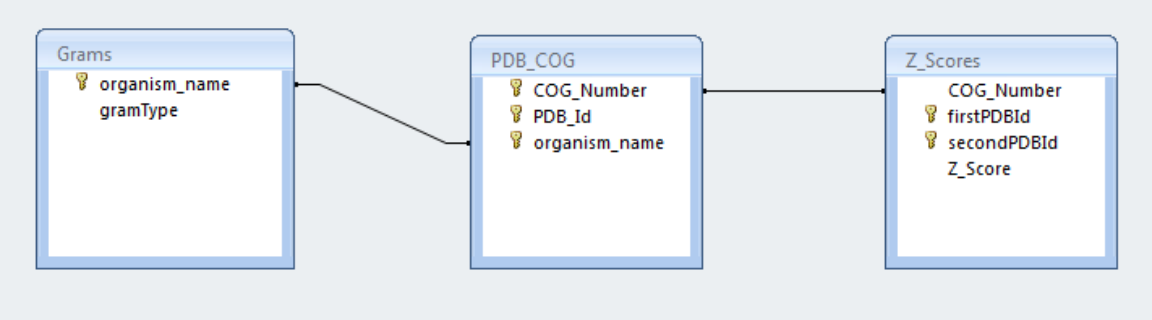


Figure 3.1: Database schema used for automation

A user interface was created in which users could chose pairs of opposite Gram bacteria, and the formatted data with average Z-scores for each COG would be displayed. All the results can be exported as comma separated format for further analysis. Flexibility for adding data to the tables has also been provided and option for uploading their own database files has been provided as well. This application can be used by future researches in this topic and could be of great use.

With the automation program we tried to compare *E. coli* with all the Firmicutes bacteria to detect possible HGTs in *E. coli* from Firmicutes bacteria. The automation process greatly reduced the data set to be analyzed by hand.

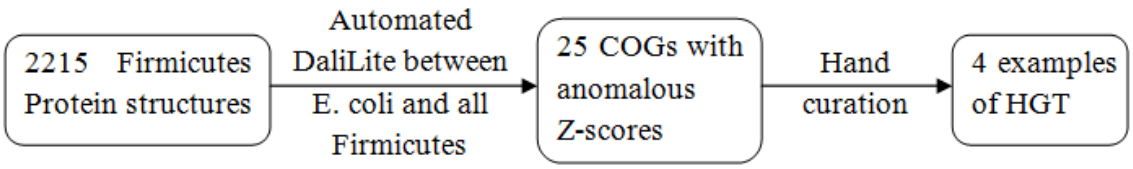


Figure 3.2: Flow chart of the method

Chapter 4

Analysis and Results

We needed to do our analysis on proteins from two bacterial phyla. *E. coli* was chosen as the representative organism from Proteobacteria because it is among the most extensively studied bacteria and has the most number of crystallized proteins.

The protein structures of *E. coli* were compared with all the Firmicutes (Gram positive) bacteria having greater than 40 of crystallized proteins in the PDB. There were 15 Gram positive organisms with crystallized proteins greater than 40. But *E. coli* could be compared to only 7 of them that had COG numbers matching with the ones *E. coli* has.

The Gram positive organisms compared with *E. coli* are:

1. *Bacillus subtilis*
2. *Staphylococcus aureus*
3. *Bacillus stearothermophilus*
4. *Streptococcus pneumonia*
5. *Lactococcus lactis*
6. *Bacillus anthracis*
7. *Bacillus megaterium*

The comparison of protein structure within the common COGs of *E. coli* and the other Gram positive organism is tabulated in tables 4.1-4.7. The highlighted row are the COGs that have the average Z-score values in that COG less than or equal to 75% of the average Z-score value in the other organism within the same COG (this is because we are looking for HGT in *E. coli* otherwise the criteria would have been the other way round). These are the COGs of our interest and a further analysis is continued only on these.

Table 4.1: Z-score structural comparison between *Escherichia coli* and *Bacillus subtilis*

COG Number	*Escherichia coli*	*Bacillus subtilis*	Comparison	*Escherichia coli* Normalized	*Bacillus subtilis* Normalized	Comparison Normalized
840	5.6	29.6	1.625	0.19	1	0.054
500	12	39.5	16.55	0.30	1	0.42
1278	5.1	13.2	7.21	0.39	1	0.55
596	24.8	47.6	28.13	0.52	1	0.59
604	36.76	55	41.1	0.67	1	0.75
789	10.35	15.2	8.33	0.68	1	0.55
1609	28.95	41.6	28.87	0.7	1	0.69
526	13.38	19.2	10.2	0.7	1	0.53
503	20.66	27.63	11.8	0.75	1	0.43
511	12.9	16.27	7.93	0.79	1	0.49
2141	53.3	63.7	32.5	0.84	1	0.51
563	31.67	37.1	29.25	0.85	1	0.79
171	40.23	42.83	36.31	0.94	1	0.85
2202	22.58	23.5	10.37	0.96	1	0.44
34	60.31	62.2	44.92	0.97	1	0.72
1985	51.97	53.1	38.17	0.98	1	0.72
363	45.07	45.7	37.39	0.99	1	0.82
207	43.21	42.13	33.62	1	0.98	0.78
236	17.35	8.9	11.75	1	0.51	0.68
454	35.7	12.09	9.71	1	0.34	0.27
653	53.1	44.3	40.01	1	0.83	0.75
745	12.55	0	9.43	1	0	0.75
784	23.7	21.48	16.567	1	0.91	0.7
834	25.7333	23.2	22.283	1	0.94	0.87
1057	33.1	23.7	22.05	1	0.72	0.67
1309	22.88	14.27	10.65	1	0.62	0.47
1925	20.85	13.05	12.9	1	0.63	0.62
2050	22.59	21.3	15.12	1	0.94	0.67
2113	49.3	41	9.13	1	0.83	0.19
2132	71.01	67.53	44.35	1	0.95	0.62
2217	14.3	12.01	10.38	1	0.84	0.73
2351	24.6	23.93	17.32	1	0.97	0.7
4948	52.3	44.77	37.68	1	0.85	0.72

Table 4.2: Z-score structural comparison between *Escherichia coli* and *Staphylococcus aureus*

COG Number	*Escherichia coli*	*Staphylococcus aureus*	Comparison	*Escherichia coli* Normalized	*Staphylococcus aureus* Normalized	Comparison Normalized
441	27.20	59.1	41.28	0.46	1	0.7
526	13.38	23.63	13.67	0.567	1	0.579
614	27.8	43.6	21.6	0.64	1	0.5
5640	31.94	44.3	31.88	0.72	1	0.72
242	32.18	33.78	21.0	0.95	1	0.62
1057	33.1	33.7	23.25	0.98	1	0.69
2367	43.41	43.97	34.87	0.99	1	0.79
24	47.82	47.97	35.82	1	1	0.75
125	38.6	34.2	24.65	1	0.89	0.64
162	50.98	47.7	42.45	1	0.94	0.83
454	35.7	9.7	8.1	1	0.27	0.23
584	55.9	47.8	29.45	1	0.86	0.53
4948	52.3	49.8	26.48	1	0.95	0.51

Table 4.3: Z-score structural comparison between *Escherichia coli* and *Bacillus stearothermophilus*

COG Number	*Escherichia coli*	*Bacillus stearothermophilus*	Comparison	*Escherichia coli* Normalized	*Bacillus stearothermophilus* Normalized	Comparison Normalized
532	0	0	2.38	0	0	1
508	3.91	7.5	4.01	0.52	1	0.54
80	6.99	13.2	5.12	0.53	1	0.39
266	27.27	39.17	22.68	0.7	1	0.58
522	20.06	27.5	15.65	0.73	1	0.57
1194	37.84	48	31.37	0.79	1	0.65
149	42.6	47.6	37.2	0.89	1	0.78
205	46.9	48.7	44.88	0.96	1	0.92
57	54.92	56.45	50.98	0.97	1	0.9
112	72.6	70.1	62.23	1	0.97	0.86
162	50.98	31.15	36.82	1	0.61	0.72
210	48.32	28.22	31.56	1	0.58	0.65
358	14.79	0	1.33	1	0	0.09
359	13.62	8.37	2.93	1	0.61	0.21
749	55.68	54.57	37.96	1	0.98	0.68
776	11.05	9.8	9.37	1	0.89	0.85
784	23.7	18.1	16.5	1	0.76	0.7
1438	16.13	15.4	10.63	1	0.95	0.66
1925	20.86	12.1	12.74	1	0.58	0.61

Table 4.4: Z-score structural comparison between *Escherichia coli* and *Streptococcus pneumonia*

COG Number	*Escherichia coli*	*Streptococcus pneumoniae*	Comparison	*Escherichia coli* Normalized	*Streptococcus pneumoniae* Normalized	Comparison Normalized
745	12.55	25.22	16.17	0.5	1	0.64
242	32.18	36.28	20.39	0.89	1	0.56
136	54.43	59.92	38.82	0.91	1	0.65
128	57.92	61.5	41.81	0.94	1	0.68
304	72.17	76.23	62.81	0.95	1	0.82
1207	49.18	50.85	42.02	0.97	1	0.83
494	13.64	13.5	11.14	1	0.99	0.82

Table 4.5: Z-score structural comparison between *Escherichia coli* and *Lactococcus lactis*

COG Number	*Escherichia coli*	*Lactococcus lactis*	Comparison	*Escherichia coli* Normalized	*Lactococcus lactis* Normalized	Comparison Normalized
266	27.27	42.3	22.74	0.64	1	0.54
2376	38.67	51.8	34.81	0.75	1	0.67
40	37.4	48.8	22.43	0.77	1	0.46

Table 4.6: Z-score structural comparison between *Escherichia coli* and *Bacillus anthracis*

COG Number	*Escherichia coli*	*Bacillus anthracis*	Comparison	*Escherichia coli* Normalized	*Bacillus anthracis* Normalized	Comparison Normalized
5126	13.53	41.06	11.79	0.33	1	0.29
329	46.76	53.1	41.95	0.88	1	0.79
605	33.58	34	31.66	0.99	1	0.93
171	40.23	36.13	33.73	1	0.9	0.84
783	29.67	26.6	19.11	1	0.9	0.64

Table 4.7: Z-score structural comparison between *Escherichia coli* and *Bacillus megaterium*

COG Number	*Escherichia coli*	*Bacillus megaterium*	Comparison	*Escherichia coli* Normalized	*Bacillus megaterium* Normalized	Comparison Normalized
1925	20.86	42.98	15.16	0.49	1	0.35
1609	28.96	41.32	27.94	0.7	1	0.68
1028	34.77	46.4	33.47	0.75	1	0.72

Table 4.8: Summary of candidates for HGT among the compared protein structures

COG	Bacterial Pairs		Findings
	No. of structures in *E. coli*	No. of structures in *Bacillus subtilis*	
500	2	2	Statistically promising example of HGT, provided there were more structures.
503	6	4	Most likely a good example of HGT.
526	38	13	Substrate diversity.
596	2	2	Most likely a good example of HGT.
604	3	2	Most likely a good example of HGT.
789	6	2	Most likely a good example of HGT. But a closer examination revealed it was the result of protein fragments in *E. coli*.
840	2	2	The two Gram-positive protein structures are not different and not similar to any of the Gram-negative protein structures.
1278	2	4	Most likely a good example of HGT.
1609	42	2	Substrate diversity.
	No. of structures in *E. coli*	No. of structures in *Staphylo-coccus aureus*	
441	9	2	Protein fragments in *E. coli* and the two Gram-positive proteins are not different
526	38	4	Substrate diversity
614	8	2	The two Gram-positive protein structures are not different and have similar Z-scores to all the protein structures in Gram-negative.
5640	15	3	The three Gram-positive protein structures are not different and have similar Z-scores to all the protein structures in Gram-negative.
	No. of structures in *E. coli*	No. of structures in *Bacillus stearother-mophilus*	
80	30	2	NULL values of Z-scores, Substrate diversity, Protein fragments*.
266	6	12	Substrate diversity, confirmation changes.

508	6	8	NULL values of Z-scores, Protein domains & fragments*.
522	33	2	Substrate diversity, NULL values of Z-scores, Protein domains/ fragments*.
	No. of structures in *E. coli*	No. of structures in *Streptococcus pneumoniae*	
745	16	4	NULL values of Z-scores, Protein domains & fragments*, same protein crystallized more than once.
	No. of structures in *E. coli*	No. of structures in *Lactococcus lactis*	
266	6	7	Conformation changes.
2376	5	2	Different subunit of a multi subunit enzyme, so the structures are unrelated but is not a HGT.
	No. of structures in *E. coli*	No. of structures in *Bacillus anthracis*	
5126	3	10	Same proteins with and without ligand, Substrate diversity, HGT not from any Gram-positive bacteria.
	No. of structures in *E. coli*	No. of structures in *Bacillus megaterium*	
1028	7	4	Substrate diversity.
1609	42	9	Substrate diversity.
1925	8	4	Protein domains & fragments*.

* In cases where protein fragments are involved, other methods can be used instead of the Z-score comparison. For example, we could use Revesz's sequence tilting method.[32] Revesz's tiling method approximately reconstructs the entire sequence of a protein using fragments of another protein. The measure of the goodness of the tiling between two strings *a* and *b*, called the tiling similarity, is defined as:

$$\mathrm{TS}(a, b) = \frac{\text{sum of the similarities in the alignments}}{\text{number of tiles in the tiling}}$$

If there are several possible tilings, we need to choose the tiling that yields the highest tiling similarity score.[32]

4.1 Summary of Suspected HGT

A further detailed analysis of all the proteins in these candidate HGTs resulted in identification of the proteins 2DY0 in COG-503, 1M33 in COG-596, 1O98 & 1O8C in COG-604 and 3MEF in COG-1278 as possible HGT to *E. coli* from *Bacillus subtilis*.

Table 4.9: Summary of Proteins suspected as HGT

PDB-ID	COG	ΔZ-score*	Receiving Bacteria	Donor Bacteria
2DY0	503	11.85	*Escherichia coli*	*Bacillus subtilis*
1M33	596	4.95	*Escherichia coli*	*Bacillus subtilis*
(1O98, 1O8C)	604	15.45	*Escherichia coli*	*Bacillus subtilis*
3MEF	1278	5.28	*Escherichia coli*	*Bacillus subtilis*

* The ΔZ-score is the difference of the average comparison Z-scores of the HGT suspected protein with all the proteins in the opposite Gram organism and the average Z-scores of all the other proteins in the same COG as the suspected protein with all the proteins in the opposite Gram organism.

4.2 Detailed Analysis of COG-503

COG-503 from *E. coli* includes five structures of *Xanthine Transferase* (1A95, 1A96, 1A97, 1A98, 1NUL) and one structure of *Adenine Transferase* (2DY0). Among these the *Adenine Transferase* had the most divergent structure according to the Z-score comparison; an average of 10 compared to an average of 25 for all the others.

COG-503 from *Bacillus subtilis* includes four structures, one *Repressor* (1O57) and 3 *Xanthine Transferase* (1P96, 1Y0B, 2FXV). All of the four proteins were closely related according to their Z-scores.

E. coli protein 2DY0 was more similar to the four *Bacillus subtilis* proteins than it was to the *E. coli* proteins. Therefore, it is an excellent candidate to be a horizontally transferred gene product. This example has not been reported in the literature.

Table 4.10: COG- 503 in Comparison between *Escherichia coli* and *Bacillus subtilis*

E. coli proteins versus each other						
	1A95	1A96	1A97	1A98	1NUL	2DY0
1A95		29.7	28.3	22.7	26	11.3
1A96			28.3	22.7	26	11.3
1A97				23.2	26.2	11
1A98					23.5	9.6
1NUL						10.2
2DY0						

Bacillus subtilis proteins versus each other				
	1O57	1P4A	1Y0B	2FXV
1O57		39.9	23	23.6
1P4A			22.9	23.6
1Y0B				32.8
2FXV				

E. coli versus *Bacillus subtilis* proteins				
	1O57	1P4A	1Y0B	2FXV
1A95	9.9	10.8	10.3	10.1
1A96	9.9	10.9	10.3	10.1
1A97	9.7	10.6	10	9.8
1A98	8.5	9.3	9.6	8.9
1NUL	9.1	9.9	9.4	9.3
2DY0	20.5	20.3	23.7	22.2

To further confirm this is a genuine case of HGT, we compared visually the 3-D structure of the protein 2DY0 and a sequence alignment with the proteins in *Bacillus subtilis* and other proteins in *E. coli* in the COG-503.

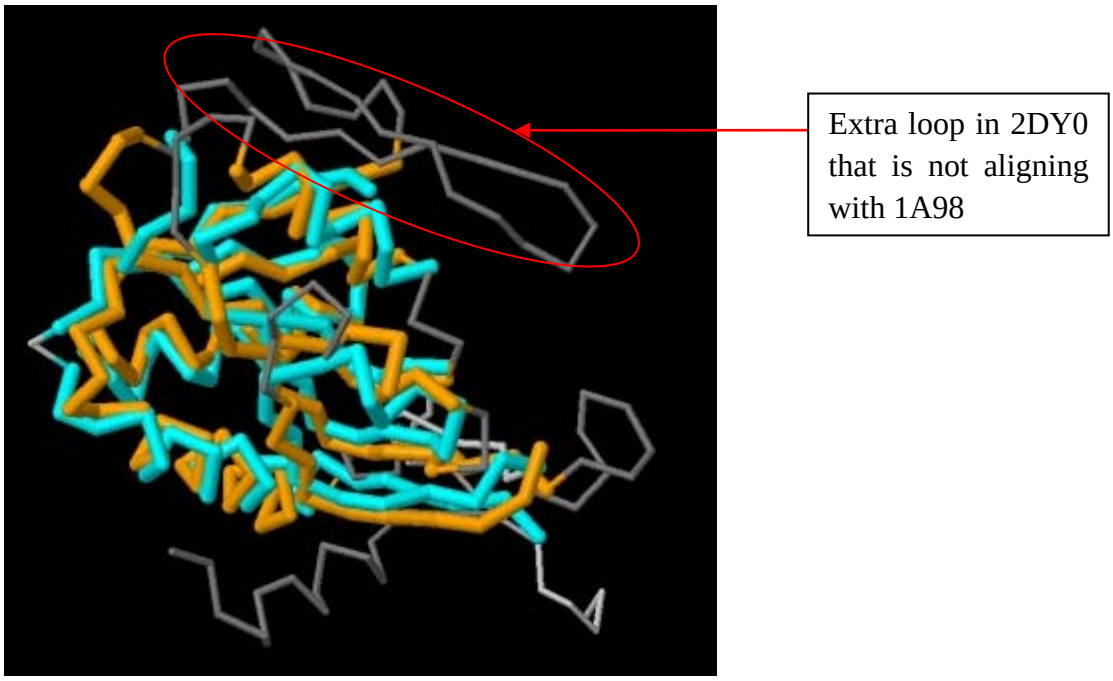


Figure 4.1: Pre-calculated jFATCAT-rigid structure alignment results 2DY0 (*E. coli*) vs. 1A98 (*E. coli*)

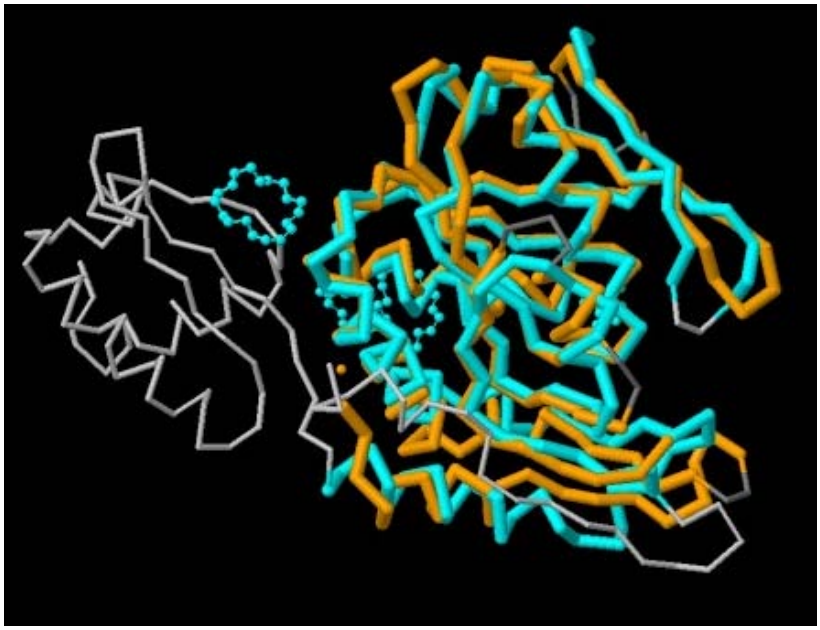


Figure 4.2: Pre-calculated jFATCAT-rigid structure alignment results 2DY0 (*E. coli*) vs. 1O57 (*Bacillus subtilis*)

Long part of the sequence is not aligning here. This corresponds to the extra grey loop in the 3-D structural comparison above.

Figure 4.3: Sequence alignment results 2DY0 (*E. coli*) vs. 1A98 (*E. coli*)

Figure 4.4: Sequence alignment results 2DY0 (*E. coli*) vs. 1O57 (*Bacillus subtilis*)

4.3 Detailed Analysis of COG-596

COG-596 from *E. coli* includes two structures one of a BioH protein (1M33) and one of a C-C bond hydrolase (1U2E).

COG-596 from *Bacillus subtili*s includes two structures of the same Sigma factor SigB regulation protein.

E. coli protein 1M33 was more similar to the protein in *Bacillus subtilis* than it had been to the *E. coli* protein. Therefore, it is an excellent candidate to be a horizontally transferred gene product.

Table 4.11: COG- 596 in Comparison between *Escherichia coli* and *Bacillus subtilis*

E. coli proteins versus each other		
	1M33	1U2E
1M33		24.8
1U2E		

Bacillus subtilis proteins versus each other		
	1WOM	1WPR
1WOM		47.6
1WPR		

E. coli versus *Bacillus subtilis* proteins		
	1WOM	1WPR
1M33	30.5	30.7
1U2E	25.7	25.6

To further confirm this is a genuine case of HGT, we compared visually, the 3-D structure of the protein 1M33 and a sequence alignment with the proteins in *Bacillus subtilis* and other proteins in *E. coli* in the COG-596.

More grey regions in this alignment of 1M33 & 1U2E compared to the alignment of 1M33 & 1WOM

Figure 4.5: Pre-calculated jFATCAT-rigid structure alignment results 1M33 (*E. coli*) vs. 1U2E (*E. coli*)

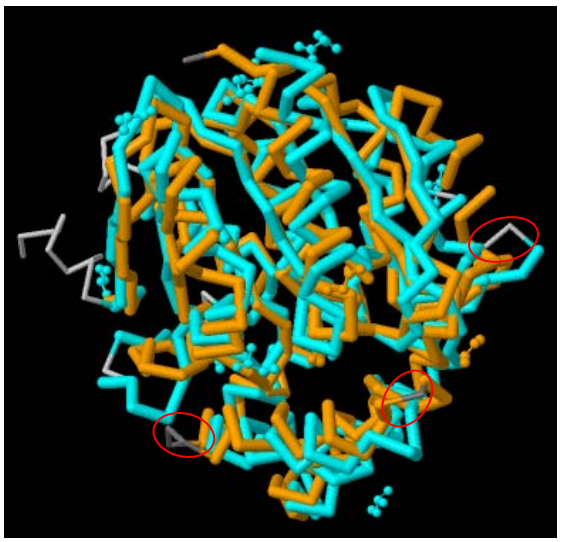
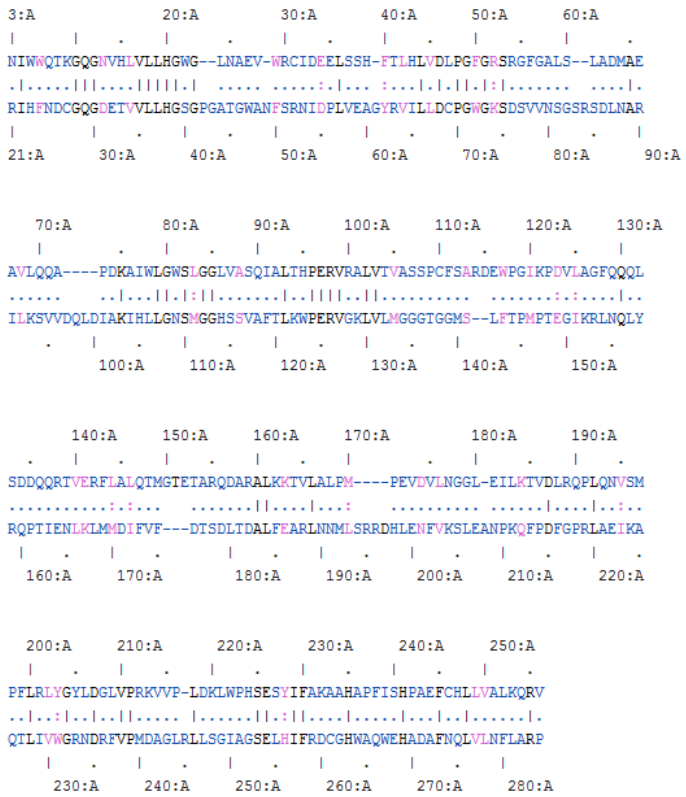


Figure 4.6: Pre-calculated jFATCAT-rigid structure alignment results 1M33 (*E. coli*) vs. 1WOM (*Bacillus subtilis*)



Sequence alignment clearing showing more gaps in the alignment of 1M33 vs. 1U2E when compared to the alignment of 1M33 vs. 1WOM

Figure 4.7: Sequence alignment results 1M33 (*E. coli*) vs. 1U2E (*E. coli*)

Figure 4.8: Sequence alignment results 1M33 (*E. coli*) vs. 1WOM (*Bacillus subtilis*)

4.4 Detailed Analysis of COG-604

COG-604 from *E. coli* includes three structures, two of which (1O89, 1O8C) are of the same protein which is an YhdH putative quinone oxidoreductase and one of a quinone oxidoreductase (1QOR).

COG-604 from *Bacillus subtilis* includes two structures of the same YhfP hypothetical protein without and with NAD bound (1TT7, 1Y9E).

E. coli protein 1O89 & 1O8C are more similar to all the structures of the protein in *Bacillus subtilis* than it had been to the *E. coli* protein 1QOR. Therefore, they are an excellent candidate to be a horizontally transferred gene product.

In this case we cannot really distinguish between the proteins 1O89, 1O8C and pin point one of them as the candidate for HGT as they are of the same protein.

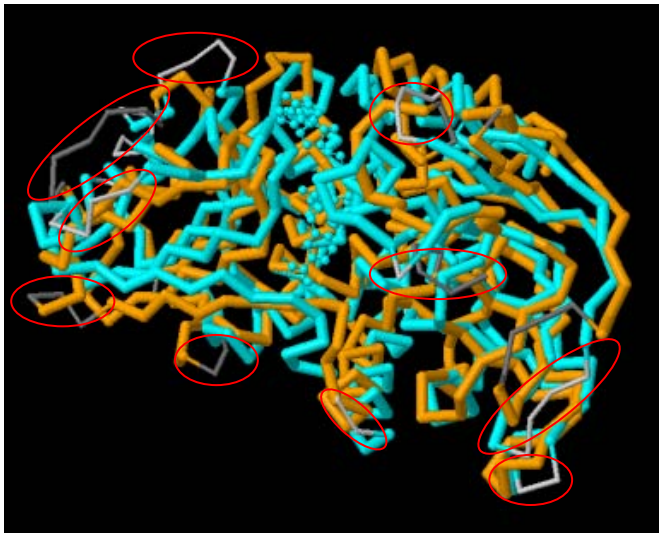
Table 4.12: COG-604 in Comparison between *Escherichia coli* and *Bacillus subtilis*

E. coli proteins versus each other			
	1O89	1O8C	1QOR
1O89		49.4	29.1
1O8C			31.8
1QOR			

Bacillus subtilis proteins versus each other		
	1TT7	1Y9E
1TT7		55
1Y9E		

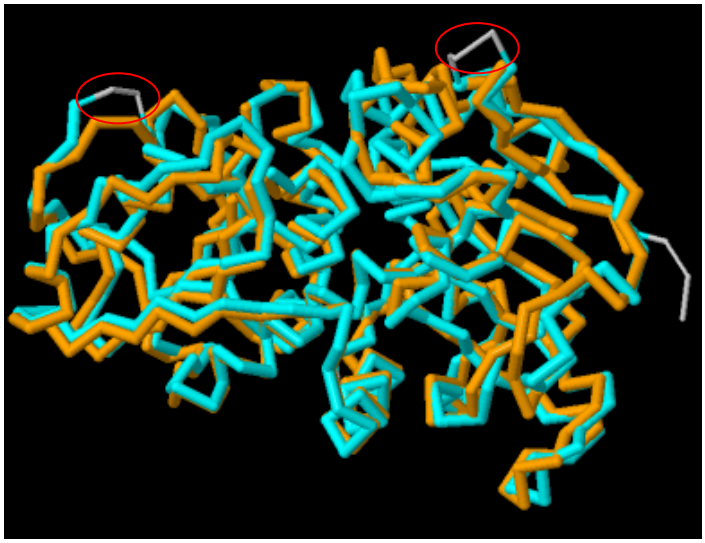
E. coli versus *Bacillus subtilis* proteins		
	1TT7	1Y9E
1O89	44.8	44.9
1O8C	47.7	47.6
1QOR	30.9	30.7

To further confirm this is a genuine case of HGT, we compared the 3-D structure of the protein 1O89 (chose one of the two similar proteins) and a sequence alignment with the proteins in *Bacillus subtilis* and other protein in *E. coli* in the COG-604.



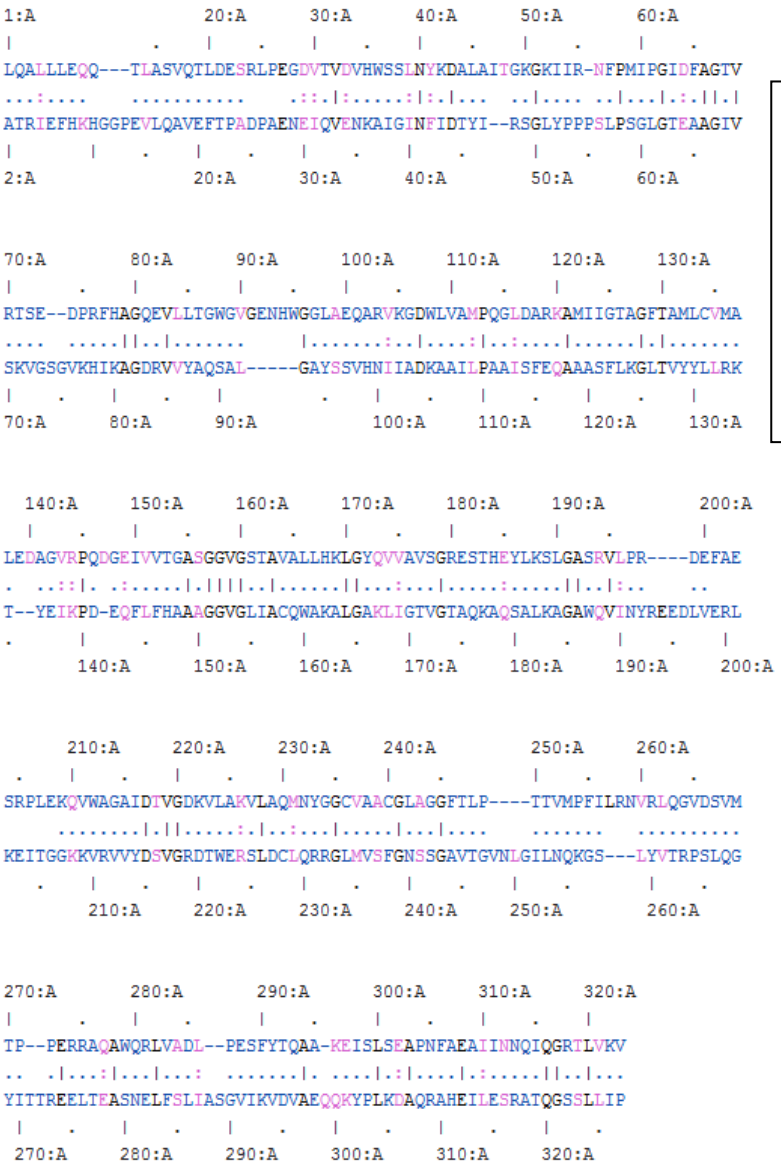
Numerous small grey loops spread throughout the alignment of 1O89 & 1QOR.

Figure 4.9: Pre-calculated jFATCAT-rigid structure alignment results 1O89 (*E. coli*) vs. 1QOR (*E. coli*)



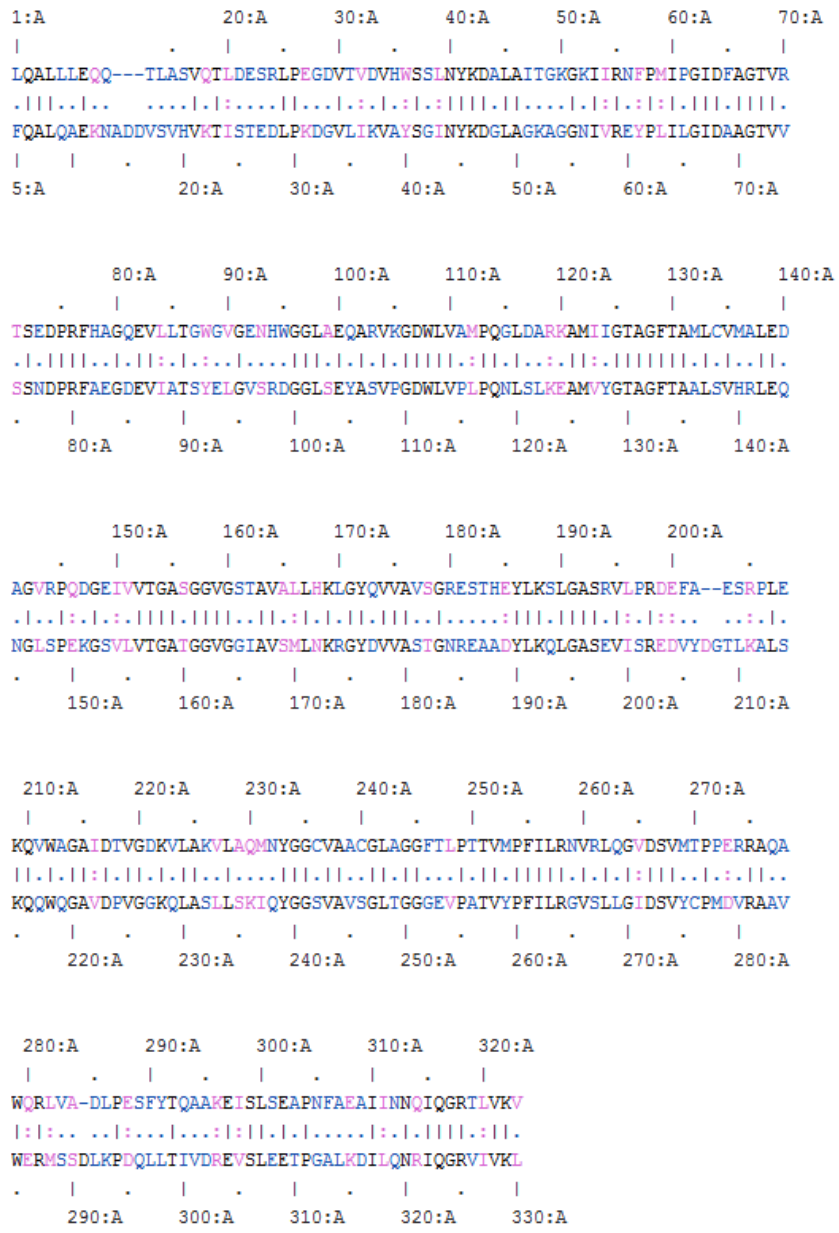
Very few loops in the alignment of 1O89 & 1TT7

Figure 4.10: Pre-calculated jFATCAT-rigid structure alignment results 1O89 (*E. coli*) vs. 1TT7 (*Bacillus subtilis*)



The sequence alignment of 1O89 & 1QOR clearly shows a lot of small mismatch which correspond to the numerous small grey regions in the 3-D structure alignment.

Figure 4.11: Sequence alignment results 1O89 (*E. coli*) vs. 1QOR (*E. coli*)



The sequence alignment of 1O89 & 1TT7 have very few mismatches.

Figure 4.12: Sequence alignment results 1O89 (*E. coli*) vs. 1TT7 (*Bacillus subtilis*)

4.5 Detailed Analysis of COG-1278

COG-1278 from *E. coli* includes two structures of two different cold shock proteins (2BH8, 3MEF).

COG-1278 from *Bacillus subtilis* includes four protein structures of the same protein, two of which are with different ligands (2ES2, 2F52) and two of which are site mutants of the same protein (2I5L, 2I5M).

E. coli protein 3MEF is more similar to all the different structures of the proteins in *Bacillus subtilis* than it had been to the *E. coli* protein 2BH8. Therefore, it is an excellent candidate to be a horizontally transferred gene product. This example has not been reported in the literature.

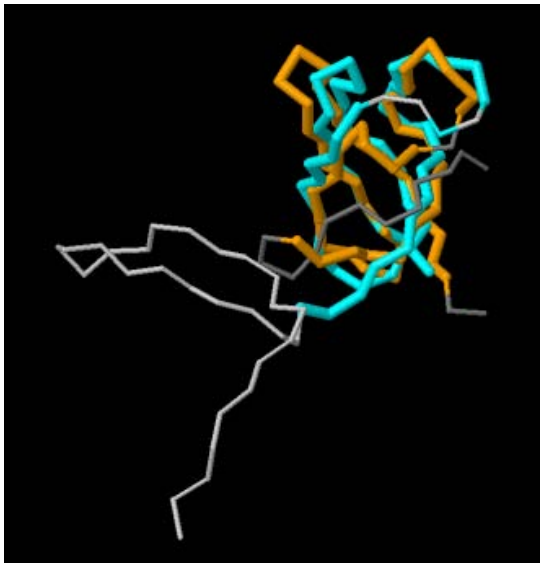
Table 4.13: COG-1278 in Comparison between *Escherichia coli* and *Bacillus subtilis*

E. coli proteins versus each other		
	2BH8	3MEF
2BH8		5.1
3MEF		

Bacillus subtilis proteins versus each other				
	2ES2	2F52	2I5L	2I5M
2ES2		12.1	15	14.6
2F52			11.7	10.8
2I5L				15
2I5M				

E. coli versus *Bacillus subtilis* proteins				
	2ES2	2F52	2I5L	2I5M
2BH8	4.7	4	4.8	4.8
3MEF	10.4	9	10.2	9.8

To further confirm this is a genuine case of HGT, we compared visually the 3-D structure of the protein 3MEF and a sequence alignment with the proteins in *Bacillus subtilis* and other proteins in *E. coli* in the COG-1278.

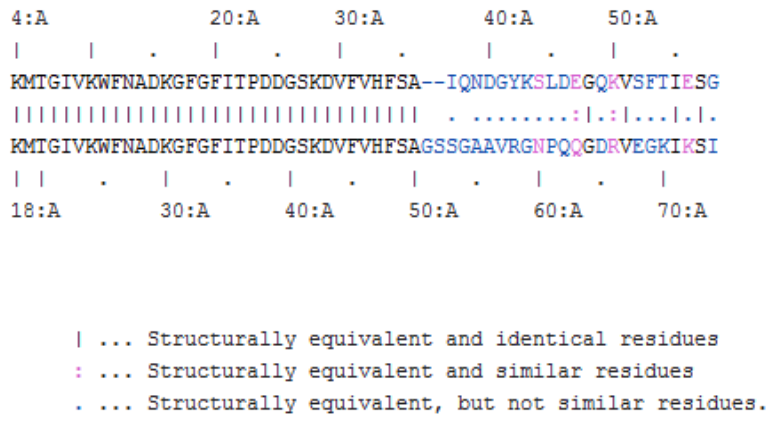


A detailed examination of the 2BH8 protein revealed possible errors in the backbone assignment but these don't alter the location of the small grey loops.

Figure 4.13: Pre-calculated jFATCAT-rigid structure alignment results 3MEF (*E. coli*) vs. 2BH8 (*E. coli*)

Structure of 3MEF better matches with 2ES2 than 2BH8.

Figure 4.14: Pre-calculated jFATCAT-rigid structure alignment results 3MEF (*E. coli*) vs. 2ES2 (*Bacillus subtilis*)



Sequence of 3MEF better matches with 2ES2 than 2BH8.

Figure 4.15: Sequence alignment results 3MEF (*E. coli*) vs. 2BH8 (*E. coli*)

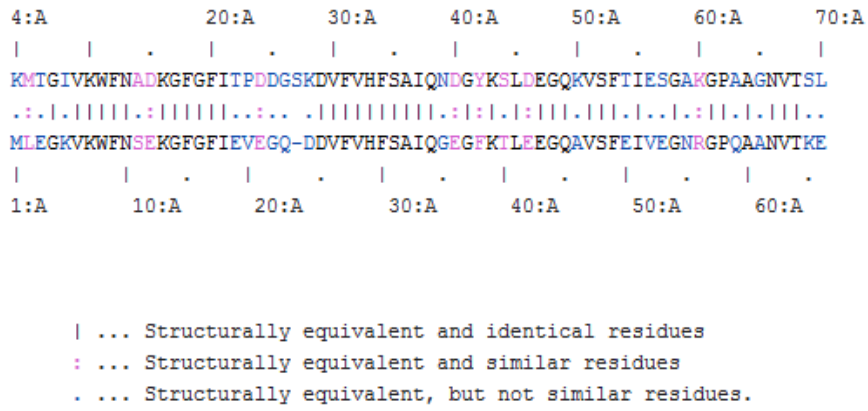


Figure 4.16: Sequence alignment results 3MEF (*E. coli*) vs. 2ES2 (*Bacillus subtilis*)

4.6 False Positives

Initially the analysis on these COGs with suspicious HGTs seemed to have found a very a large number of HGTs. However, an intensive analysis proved that many of these were false positives. There were the following reasons for false positives:

1. **Protein Fragments**: Many of the PDB-ids in the Protein Data Bank correspond to Protein domains and Protein fragments. The structural comparison of these Domains and Protein fragments with the whole protein sometimes leads to falsely suspecting a protein for HGT.

   Good examples of this case are COG-1925 and COG-2376.
2. **Substrate Diversity**: The COG's enzyme specificity is fixed within the COG but the substrate specificity is diverse.

   Good examples for this case are COG-526 and COG-1609.
3. **Conformation changes**: There are two or more conformations of the same protein. Example: COG-266
4. **NULL values**: Comparison of structures with no significant similarity should be considered a 'NULL'. This disturbs the statistical analysis greatly.
5. **HGT from other sources**: There are some cases in which a protein is identified as possible HGT but not exactly from the organism with which we are comparing. Example: Protein 1BJF in COG-5126.
6. **Different Subunits**: Different subunits of a multi subunit enzyme have very dissimilar structures and with the structure-based method these could look like a possible candidate of HGT but they are not.

Chapter 5

Relative COG Functional Similarity Based on GO Terms

5.1 Introduction

This chapter describes the work done for the paper "Bacterial Protein Structures Reveal Phylum Dependent Divergence" by Matthew D. Shortridge, Thomas Triplet, Peter Revesz, Mark A. Griep, and Robert Powers, Computational Biology and Chemistry, accepted, January 2011.[32]

PDB is generally used when working with protein structures. There are various types of functional classifications of the proteins. The mains ones used are the COG classification and the GO annotation. The mapping of PDB IDs and COG classification has been done at UNL and available on PROFESS database. The GO Annotation project has mapped PDB IDs and corresponding chain IDs to the GO terms.

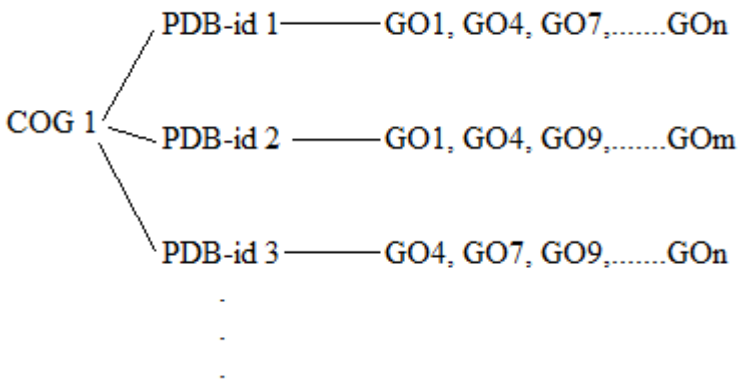
In most of the cases the proteins in the same COG classification have structures similar to each other. But many a times in research we would like to know which COGs have

proteins more related to each other than the other based on the functions of the proteins. That is where the GO term classification comes into picture where each protein has a few GO terms related to it corresponding to the various functions of the protein.

So the data that one would be working on would look like this:

COG1 = {PDBid1, PDBid2, ... PDBidn}

PDBid1 = {GO1, GO2, ... GOi}, PDBid2 = {GO1, GO4, ... GOj}, ... PDBidk = {GO1, GO9, ... GOk}



5.1.1 Hamiltonian Distance

We use the Hamiltonian distances between sets as the basis for determining the functional similarity in a COG.

Hamiltonian distance between two sets can be defined as the sum of the number of values in one set that do not appear in the other and vice versa.

Example: Set A = {11, 278, 999, 1122}, Set B={11,131,278,777}

The Hamiltonian distance between Set A and Set B is 4. 999 and 1122 are not present in Set B. That makes the count 2. Then 131 and 777 are not present in Set A and adds 2 more to our count making the total count 4 which is the distance.

5.2 Method

By definition, a strong consensus requires each protein to share the same GO term. Instead, we define weak consensus as a set of GO terms that appear in a majority of proteins. We can have different thresholds according to our requirement. We chose a threshold of 50% in our experiments. So the weak consensus set for COG1 in the example above

Weak consensus of COG1 ={GO1, GO4, GO7, ... GOk}

Where each GOi appears in more than 50% of the total number of proteins in COG1.

To measure the similarity of the proteins based on GO terms with in a COG we can adopt two different methods. One using the weak consensus set and one without considering the weak consensus set.

5.2.1 Method-1

In this, we first get the weak consensus set of GO terms for the COG of interest, WC. Then, we find the distance between the WC set and GO term set for each Protein in the COG, normalize them by dividing the distance by the number of unique GO terms in the two sets. We then sum all the normalized distances for each of the proteins and average it

by dividing the sum with the number of proteins in that COG. The reason we normalize is to keep our distance result between 0 and 1. This makes it possible to compare with the results of other COGs.

The set function formula would look like:

$$COG_{dist\ 1} = \sum \frac{|WC\ \cup\ GO_i| - |WC\ \cap GO_i|}{|WC\ \cup\ GO_i|} \Big/ N$$

WC is the weak consensus set and GO is the GO term sets.

5.2.2 Method-2

In the second method we do not use the weak consensus set at all. Instead we do comparison between all the Proteins' GO term sets. That is, we find the distance between each and every GO term set for the proteins in a COG. Normalize each of the distances by dividing the distance by the number of unique proteins in the two comparing GO term sets considered together and then average this normalized distance by the total number of comparisons.

The set function formula would look like

$$COG_{dist\ 2} = 2\ . \sum_i \sum_{j=i+1} \frac{|GO_i\ \cup\ GO_j| - |GO_i\ \cap\ GO_j|}{|GO_i\ \cup\ GO_j|} \Big/ N(N-1)$$

Both the methods measure how functionally similar the COGs are. There is no criterion for selecting one method over the other. Actually, there are some programs like FunSimMat - Functional Similarity Matrix available that does similar comparisons.

5.3 Results

The results using this program using Method 1 and Method 2 for various COGs are as below. For both methods we calculate 1- COG_{dist} to get 1 for perfect similarly and 0 for absolute dissimilarity.

Table 5.1: Results of COG functional similarity with Method 1 and Method 2

COG	COG Function Annotation	Relative COG Function Similarity Based on GO (Method 1)	Relative COG Function Similarity Based on GO (Method 2)
28	Thiamine pyrophosphate requiring enzymes	0.59	0.40
39	Malate/lactate dehydrogenases	0.8	0.68
394	Protein-tyrosine-phosphatase	0.61	0.52
604	NADPH:quinone reductase and related Zn-dependent oxidoreductases	0.88	0.77
605	Superoxide dismutase	0.76	0.60
742	N6-adenine-specific methylase	0.73	0.59
813	Purine-nucleoside phosphorylase	0.87	0.75
1012	NAD-dependent aldehyde dehydrogenases	0.58	0.38

1075	Predicted acetyltransferases and hydrolases with the alpha/beta hydrolase fold	0.7	0.5
1607	Acyl-CoA hydrolase	0.8	0.75
1940	Transcriptional regulator/sugar kinase	0.31	0.1
2124	Cytochrome P450	0.8	0.65
2188	Transcriptional regulators	0.89	0.8
446	Uncharacterized NAD (FAD) - dependent dehydrogenases	0.85	0.71
1057	Nicotinic acid mononucleotide adenylyltransferase	0.95	0.91
242	N-formylmethionyl-tRNA deformylase	0.87	0.75
1052	Lactate dehydrogenase and related dehydrogenases	0.89	0.84
2141	Coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases	0.76	0.65
3832	Uncharacterized conserved protein	1	1
110	Acetyltransferase (isoleucine patch superfamily)	0.56	0.34
171	NAD synthase	0.85	0.73
251	Putative translation initiation inhibitor, yjgF family	0	0
346	Lactoylglutathione lyase and related lyases	0.11	0.2
366	Glycosidases	0.51	0.46
454	Histone acetyltransferase HPA2 and related acetyltransferases	0.83	0.74
491	Zn-dependent hydrolases, including glyoxylases	0.5	0.48
500	SAM-dependent	0.59	0.37

	methyltransferases		
526	Thiol-disulfide isomerase and thioredoxins	0.96	0.93
590	Cytosine/adenosine deaminases	0.7	0.54
637	Predicted phosphatase/phosphohexomutase	0.52	0.33
664	cAMP-binding proteins	0.5	0.34
745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	0.73	0.55
753	Catalase	0.93	0.89
778	Nitroreductase	0.64	0.53
784	FOG: CheY-like receiver	0.48	0.4
796	Glutamate racemase	0.92	0.84
1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	0.84	0.71
1151	6Fe-6S prismane cluster-containing protein	0.71	0.6
1309	Transcriptional regulator	0.8	0.7
1396	Predicted transcriptional regulators	0.54	0.4
1404	Subtilisin-like serine proteases	0.6	0.5
1733	Predicted transcriptional regulators	1	1
1846	Transcriptional regulators	0.85	0.73
2159	Predicted metal-dependent hydrolase of the TIM-barrel fold	0.83	0.78
2367	Beta-lactamase class A	0.93	0.87
2730	Endoglucanase	0.88	0.78
3693	Beta-1,4-xylanase	0.89	0.8
4948	L-alanine-DL-glutamate epimerase and related enzymes of enolase superfamily	0.71	0.57

Method 1 gives higher similarity scores compared to Method 2 because Method 1 uses the weak consensus set. By the definition of the weak consensus set, the average distances of the proteins with their weak consensus sets is always less than the distance between themselves.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Identifying HGTs is a difficult process. No process or method proposed so far is capable to identify perfectly all cases of HGTs.[12] Each process has its own advantages and disadvantages. This research devised a novel protein structure-based method for identifying HGTs and has proved that working directly with the proteins and their structures is a good option and an innovative approach for identifying HGTs. The various possibilities of false positives also have been studied and documented.

A paper based on this thesis work has been submitted to the Fourth International C* Conference on Computer Science & Software Engineering (C^3S^2E, Montreal, QC, CANADA).[31]

6.2 Future Work

The process of identifying HGTs using whole organism protein structures is the first of its kind and has a vast scope for improvements and advancements. In particular, ways to eliminate each one of the cases for false positives discussed in Chapter 4 would be the highest priority for improving our method.

PDB is the best database available for the various crystallized proteins, their structures etc. However, some of the problems encountered when using PDB are:

1. There is some redundancy in the PDB i.e. some proteins that have been crystallized more than once and each appear with a unique PDB-id.
2. Some proteins have been crystallized with and without ligands and substrates, each appear with a unique PDB-id.
3. Protein Domains and Protein fragments appear with unique PDB-id.
4. Some proteins have been mutated at only one or a few residues, but each structure has a unique PDB-id.

These issues cause considerable deviation in the analysis as well as the results. Some of the false positive cases can be eliminated when the PDB gets cleaned.

There are millions of proteins in various organisms. Not all the proteins have been crystallized and their structures are not available. This is one of the main limitations of using the protein structure based approach for identifying HGT. As more and more protein structures are crystallized and as the PDB expands, the efficiency of this protein structure-based method for detecting HGT will surely get better.

Instead of the Dali method of protein comparison used in this research, we could try using the Comparison of Protein Active Site Structures[26] and see if it gives different type of results.

COG classification is more of a generalized classification of proteins and there are various other protein classifications that can be used instead of the COG. Some of the fairly recent classification like GO classification, eggNOG classification [16] etc. would be a good choice to experiment this process on. The results of the same process with a different classification could give better and more interesting results.

This research has a great potential for scalability. As more analysis is done with the other organisms and as we find more cases of HGT it would be very interesting to look into the statistics. This could include, which organism has higher percentage of HGT proteins? Which type of protein has higher cases of a HGT, etc? For all these cases we could look into the reason and this might drive us into very interesting causes and reasons. The statistics of this method could be compared to the statistics of the other methods for detecting HGT, but these statistics might not match each other because each method works in a different way.

References

1. Akiba T, Koyama K, Ishiki Y, Kimura S, Fukushima T (April 1960). "On the mechanism of the development of multiple-drug-resistant clones of Shigella". *Jpn. J. Microbiol.* **4**: 219–27.

2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). "Basic local alignment search tool". J Mol Biol 215

3. "Bacteria (eubacteria)". *Taxonomy Browser*. NCBI. Retrieved 2008-09-10.

4. Barlow M (2009). "What antimicrobial resistance has taught us about horizontal gene transfer". *Methods in Molecular Biology (Clifton, N.J.)* **532**: 397–411.

5. Bergey, David H.; John G. Holt; Noel R. Krieg; Peter H.A. Sneath (1994). *Bergey's Manual of Determinative Bacteriology* (9th ed.). Lippincott Williams & Wilkins.

6. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. Nucleic Acids Res 28, 235{42 (2000).

7. Consortium, T. G. O. The Gene Ontology (GO) project in 2006. Nucleic Acids Research 34, D322{D326 (2006).

8. Cortez D, Delaye L, Lazcano A, Becerra A. Composition-based methods to identify horizontal gene transfer. Methods Mol Biol. 2009; 532:215-25. PMID: 19271187

9. Creighton, Thomas H. (1993). "Chapter 1". *Proteins: structures and molecular properties*. San Francisco: W. H. Freeman.

10. DaliLite workbench for protein structure comparison.(Jun-2000) Bioinformatics (Oxford, England), 16 (6) :566-7

11. Garcia-Vallve S., Guzman E., Montero M.A. and Romeu A. (September 2002). "HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes".

12. Garcia-Vallve, S., Romeu, A. & Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res 10, 1719–1725.

13. Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, and Gelbart WM (2000). An Introduction to Genetic Analysis. W.H. Freeman and Company. Chapter 10 (Molecular Biology of Gene Function): Genetic code: Stop codons.

14. Holm,L. and Sander,C. (1998) Dictionary of recurrent domains in protein structures. Proteins, 33, 88-96.

15. Horizontal Gene Transfer Database, Biochemistry and Biotechnology Department, Universitat Rovira i Virgili,Taragonna, Spain.

16. Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 2008;36:D250–D254.

17. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

18. Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T (October 2002). "Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect". Proc. Natl. Acad. Sci. U.S.A. 99 (22): 14280–5.

19. La Scola B., Desnues C., Pagnier I., Robert C., Barrassi L., Fournous G., Merchat M., Suzan-Monti M., Forterre P., Koonin E., Raoult D. (September 2008). "The virophage as a unique parasite of the giant mimivirus".

20. Live on earth, Live journal. Microbiology: the first two lectures.

21. Maxmen, A. (2010). "Virus-like particles speed bacterial evolution". *Nature*. doi:10.1038/news.2010.507

22. Michael Syvanen, Cross-species Gene Transfer; Implications for a New Theory of Evolution. Harvard Medical School, Boston, MA 1984

23. Ochiai K, Yamanaka T, Kimura K, Sawada, O (1959). "Inheritance of drug resistance (and its transfer) between Shigella strains and Between Shigella and E. coli strains" (in Japanese).*Hihon Iji Shimpor* **1861**: 34.

24. Pearson H (August 2008). "'Virophage' suggests viruses are alive". *Nature* **454** (7205): 677.

25. Peter Gogarten. “Horizontal Gene Transfer - A New Paradigm for Biology” Esalan Center for theory and research

26. Powers R., Copeland J., Germer K., Mercier K., Ramanathan V., Revesz P. Comparison of Protein Active-Site Structures for Functional Annotation of Proteins and Drug Design. Proteins: Structure, Function, and Bioinformatics, vol. 65, no. 1, pp. 124-135, 2006.

27. Revesz, P. Introduction to Constraint Databases (Springer-Verlag, 2002).

28. Revesz, P. Introduction to databases: From Biological to spatio-temporal (Springer, 2010).

29. Ridley, M. (2006). *Genome*. New York, NY: Harper Perennial

30. Salton MJR, Kim KS (1996). *Structure* in: *Baron's Medical Microbiology* (Baron S *et al.*, eds.) (4th ed.). Univ of Texas Medical Branch.

31. Santosh V.R., Griep M., Revesz P. Identifying Horizontal Gene Transfer Using Anomalies In Protein Structures And Sequences. C* Conference on Computer Science & Software Engineering, submitted February 2011.

32. Shortridge M., Triplet T., Revesz P., Mark A. Griep, and Powers R. Bacterial Protein Structures Reveal Phylum Dependent Divergence. Computational Biology and Chemistry, accepted January 2011.

33. Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41 (2003).

34. Than C, Ruths D, Innan H, Nakhleh L: Identifiability issues in phylogeny-based detection of horizontal gene transfer. Comparative Genomics, Proceedings 2006, 4205:215-229.

35. Triplet T., Shortridge M., Griep M., Stark J., Powers, R. & Revesz, P. PROFESS: A PROtein Function, Evolution, Structure and Sequence database. Database (submitted) (2009).

36. Xintao Wei, Lenore Cowen, Carla Brodley, Arthur Brady, D. Sculley, and Donna K. Slonim.(2008) A Distance-Based Method for Detecting Horizontal Gene Transfer in Whole Genomes.